NASA Climate Analytics Support Biological Research

Every few seconds, NASA's satellites transmit more than a gigabyte of data back to Earth. By 2030, it's estimated the Agency will have amassed 350 petabytes of climate-change data alone. In consumer speak, consider that a single petabyte is roughly equal to the digital storage space taken up by 100,000 high-definition movies.

As NASA continues to accrue more satellite data, so do various programs that utilize them. One of those is MERRA, shorthand for Modern-Era Retrospective Analysis for Research and Applications. Run out of Goddard Space Flight Center's Global Modeling and Assimilation Office since its inception in 2008, MERRA integrates data from a variety of satellite systems into numerical models to recreate a synthetic data record of the weather.

MERRA so far has generated about 200 terabytes' worth of what climate scientists call reanalysis data, with another 2 or 3 terabytes being added each day.

"The models attempt to capture the physics that describe the atmosphere," explains Goddard senior computer scientist John Schnase. He says about 120 observation types, including satellites, radiosondes, aircraft, balloons, ship and ocean buoys, and land surface observations, provide input for MERRA. "MERRA provides global coverage. It takes the whole world and divides it up into grids and layers from the surface all the way through the stratosphere and then computes the physics going on—the heat transfers and the pressures and the moisture content."

Through this reanalysis, MERRA outputs 400 variables roughly every six hours. And it provides those results on either a worldwide or a regional level for any given timeframe from the beginning of the satellite era, in 1979, up to today.

The utility of all that data cannot be overestimated. Suppose a researcher wanted to figure out why corn
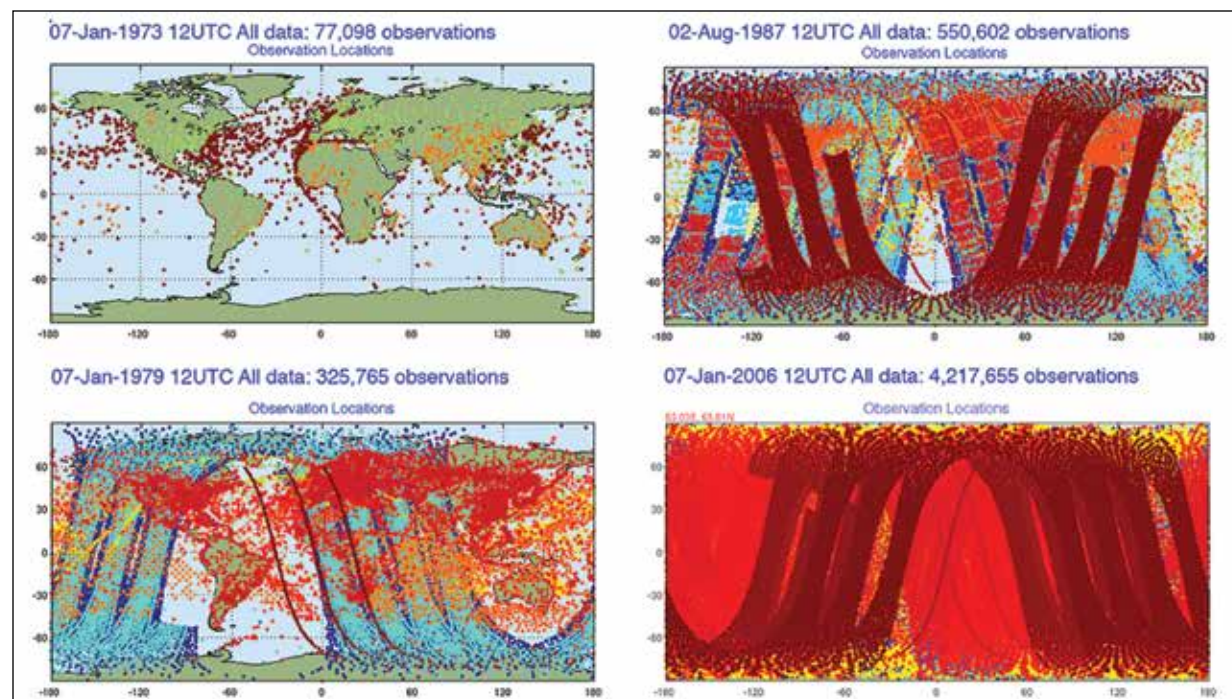
As the pictographs show, throughout the last several decades NASA has been accruing an ever-increasing amount of observational data. Run out of Goddard Space Flight Center's Global Modeling and Assimilation Office since its inception in 2008, Modern-Era Retrospective Analysis for Research and Applications, or MERRA, can integrate the data to not only recreate a synthetic record of the weather dating back to 1979 but also compute climate projections.

yields in Nebraska were so low in 1983, Schnase posits. He or she might want to use MERRA data to examine various environmental conditions during that time, such as wind speed, temperature, and humidity. "It could turn into a very challenging data-wrangling exercise in order to come up with those values independently," he says, "but MERRA provides a single place to go to find those values quickly."

Beyond looking at past weather conditions, climate models are adept at computing climate projections, which are playing an increasingly important role as Government and industry attempt to prepare for a quickly and dramatically changing planet. Scientists can tweak certain variables to see how, say, temperature swings or moisture content will affect Earth 10, 20, or 100 years out. "Reanalyses like MERRA smooth out the differences we see in various measurement systems and give us physically consistent data across all climate variables," Schnase says. "They let us calibrate forecasting systems to improve our climate projections, and they give us a unique type of data important to many other applications."

Technology Transfer

Because of their versatility, reanalyses are in high demand by outside researchers. MERRA, for example, is made available through Goddard's Earth Sciences

and Information Services Center, where people have traditionally accessed them by downloading files. And oftentimes, the file sizes are large. Really large. "You'll find that people using these big datasets are spending days, if not weeks, moving data over to their workstations," Schnase says. "That's not including the time it takes them to manipulate the data and run their experiments once they have it."

That's the crux of the big-data problem: ever-growing datasets make for better research, but it's becoming harder and harder to transfer and manipulate them.

Motivated to find a new approach to the issue, Schnase and fellow Goddard computer scientist Dan Duffy submitted a proposal in 2012 under the Agency's Research Opportunities in Space and Earth Science (ROSES) program to develop a version of the increasingly popular software-as-a-service model, in which customers access applications housed on a host vendor's servers, which perform the necessary computing work and software maintenance. Businesses of every kind utilize the model for

In the summer of 1993, massive flooding in the Midwest wreaked havoc over the region, causing $21 billion in damages and claiming 48 lives. Produced by MERRA, this information graphic shows pulses of flood-inducing moist air moving into the American Midwest during that time. Arrows indicate wind trajectories, while the various colors indicate wind altitude (black represents the lowest, white the middle, and blue the highest altitudes), and line length signifies wind speed (the longer the line, the greater the speed). Black arrows are the most telling, as low-altitude winds are responsible for carrying moisture.

Web site hosting, graphic arts programs, and IT services, among other applications.

After receiving ROSES funding in 2012, over the next few years, Schnase, Duffy, and other colleagues at Goddard's Office of Computational and Information Sciences and Technology worked to build a MERRA analytic service where outside users could employ Goddard's high-performance computing capabilities to create useful data products. As they did so, they broke the MERRA dataset up into an array of files and stored them on a storage cluster—36 servers in an arrangement that facilitates parallel computing, allowing for vast amounts of climate data to be operated on simultaneously.

For the processing program, Schnase and his team adapted an open-source program called Hadoop MapReduce, originally developed by Google as a retrieval system that allows for those near-instant search results on the tech giant's Web site. The team modified it so it would work with binary scientific data rather than text.

After those tasks were completed, and after a series of basic operations was programmed into the system, Climate Analytics-as-a-Service, or CAaaS, was born. The program was made available for beta testing in fall of 2014 through NASA's Climate Data Services application programming interface (API), which is, in essence, a programming library and set of instructions that tell software developers how to access the service. They can also use the API to design graphical interfaces that make the service user-friendly.

One of the first organizations to utilize the new technology is the iPlant Collaborative, which was established by the National Science Foundation in 2008 to develop cyberinfrastructure, including supercomputing access, to support life-sciences research. By working with NASA's API, iPlant developed a graphical interface for CAaaS that's now available to member scientists in the organization's Discovery Environment platform, a fitting name considering it houses other big data research enterprises, such as genomic sequencing. Discovery Environment currently hosts three CAaaS-supported applications: Analyzed State, Meteorology Instantaneous Monthly, and Land-Related Surface Quantities. The first two are geared for direct climate studies, and the third focuses on
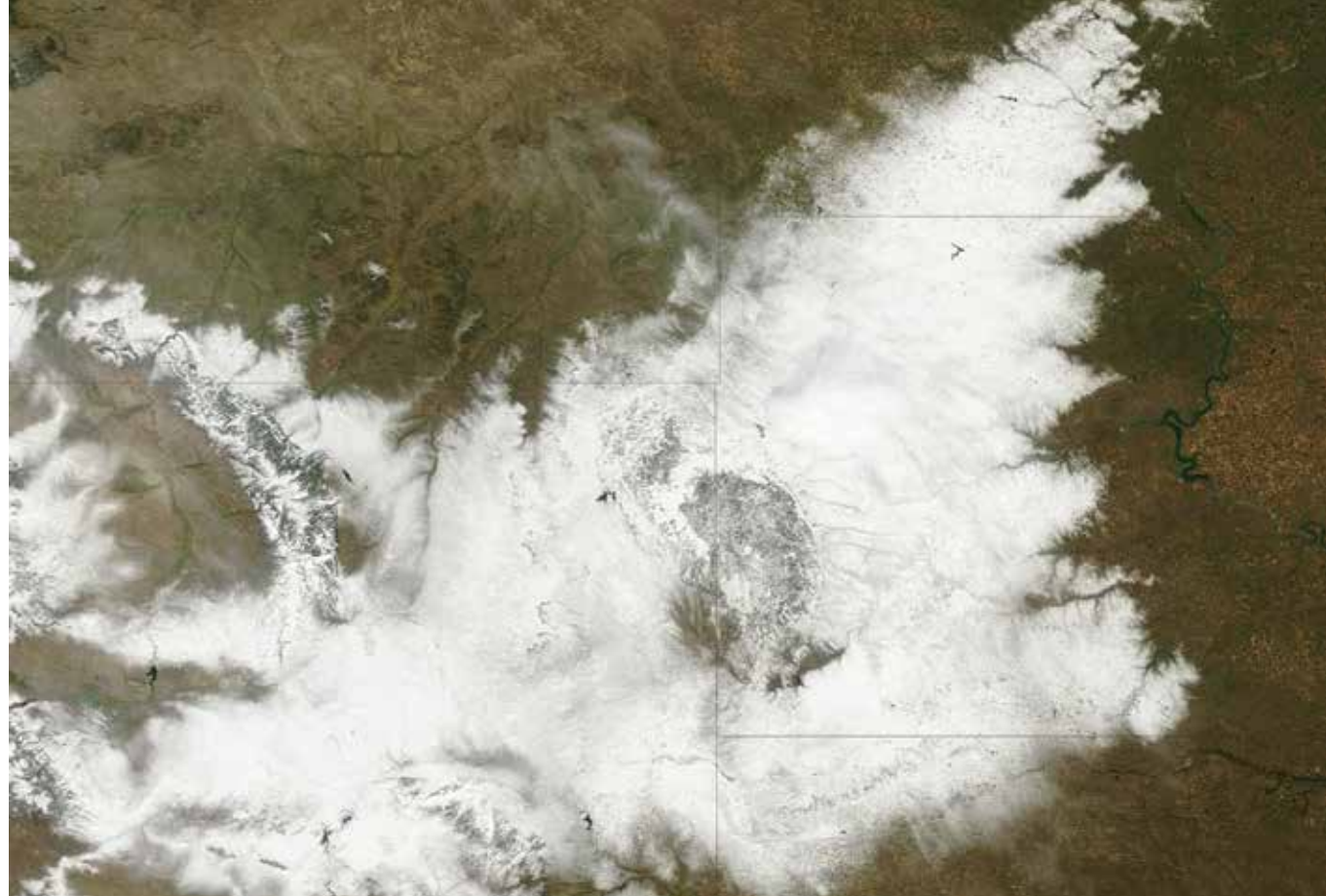
NASA's Moderate Resolution Imaging Spectroradiometer captured an image of an uncommonly early blizzard that blanketed northeastern Wyoming and western South Dakota from October 3-5, 2013. The cattle hadn't yet grown their thick winter coats that allow them to survive winters, resulting in the deaths of about 15 to 20 percent of South Dakota's herds. By using the Agency's Climate Analytics-as-a-Service, or CAaaS, researchers could run projections to determine what regions may no longer be suitable for raising cattle in the future due to climate change.

the weather's impacts on terrain, such as snow cover, soil moisture, temperature, and vegetation.

iPlant senior coordinator Martha Narro says the programs will be an important tool for biological scientists, as "One of the grand challenges is understanding how the interplay between an organism's genetic repertoire and environmental factors results in particular characteristics being expressed in the organism."

Benefits

Currently, CAaaS is set up so that MERRA can be called on to run commonly used operations for determining the minimum, maximum, or average values for variables such as temperature, humidity, wind speed, rain and snow precipitation, and evaporation. Calculations can be made on global and regional scales and can account for either a static point in time or a specified duration.

What's key is that all of the storage and computing occurs on NASA's end. Rather than moving large amounts of data to their workstations and perform these operations themselves, users only have to move the results—and those products are much smaller. CAaaS can reduce the data-wrangling time from days and weeks to minutes and hours.

"MERRA's engine does the extraction for these basic operations so they're ready to be put into more high-order analyses by the user," Schnase says, while adding that future applications using CAaaS may also contain more complex analytic functions. "One of the things we're trying to do is get scientists, nonscientists, and the private sector all using this data and making the system more robust by contributing to the development of the API."

That's just what iPlant has done by introducing visualization tools for its Discovery Environment. For example, temperatures across a region can be clearly displayed through color-coded variations on a map, Narro says. "Data like this is often stacked in layers of interrelated spreadsheets, so they are much easier to wrap your head around if you can visualize them."

Paired with those new tools, Narro envisions iPlant's applications as being able to expand research opportunities for members whose work involves planning for climate change. For one, scientists could perform crop simulation models to project where plants will grow best, given different weather conditions. Other studies may look at how projected climate changes would impact the health of livestock. "If your cattle are suffering from high temperatures, it can impact their productivity and growth," she says.

Schnase sees CAaaS being utilized in yet other ways, particularly as more types of climate model data are made available. Insurance companies will eventually be able to utilize CAaaS to fine-tune risk assessments for potential natural disasters, and urban planners can prepare cities for the impacts of climate change by developing infrastructure to counteract rising sea levels and changing weather patterns.

"This is a different approach for accessing these huge datasets, and I think it's in line with the paradigm shift that's happening in terms of how big data challenges are being addressed," Schnase says. "We're excited to see how this takes off." ❖

CAaaS

[Climate Analytics-as-a-Service]

can reduce the **data-wrangling** time from **days** and **weeks** to **minutes** and **hours.**

In this image shared by NASA's Earth Observatory, Minnesota's farmlands look like a patchwork quilt. By utilizing CAaaS, researchers can run climate projections to investigate where crops would grow best, given different scenarios.